Posture Pattern Recognition Analysis in Lectures

Weverson da Silva Pereira*, Fernando Pujaico Rivera, Leila Cristina C. Bergamasco, Paulo Sergio Silva Rodrigues
Centro Universitário FEI, São Bernardo do Campo, Brazil
wpereira@fei.edu.br*

***Abstract*—The Study of Posture Analysis and Non-Verbal Communication plays a pivotal role in enhancing communication among individuals in various contexts. The ability to decode and comprehend messages conveyed through gestures, facial expressions, and body movements is crucial for fostering more effective and meaningful interactions. Accordingly, this present work aims to conduct an exploratory analysis of posture patterns among speakers worldwide. To achieve this, the Openpifpaf algorithm was employed in videos of lectures for pose extraction, and the K-means clustering algorithm was utilized to distinguish commonly adopted postures during this lectures. The evaluation regarding the representativeness of keyposes involved an online questionnaire in which participants were asked to classify certain speaker poses into one of the clusters. The results revealed that the K-means algorithm achieved an accuracy rate of 85.71%.**

***Index Terms*—Clustering, Openpifpaf, Lectures, Pose**

I. Introduction

Body language is fundamental for social interactions, serving not only as a complement to verbal communication but also as a means to convey trust and credibility regarding the spoken word [1]. This form of communication is essential not only in interpersonal interactions but also in virtual agents that need to enhance their social interaction skills and convey emotions effectively [2]. The analysis of movement and posture patterns is a research topic intrinsically linked to the understanding of human behavior and, therefore, has been extensively studied by the scientific community.

The work conducted in [3], for instance, introduces a Computational Virtual Reality (C-VR) system capable of capturing human motion and transferring it to an avatar using inverse kinematics. Following the transfer, recordings of the character's animation are made to be employed in the avatar's pose detection process. Lastly, the authors utilized a Support Vector Machine (SVM) for pose classification and a Temporal Convolutional Network (TCN) for sequential pose analysis, both of which demonstrated superior performance to recurrent networks in sequence modeling tasks.

For the same purpose, the authors in [4] proposed a method called Pose2Pose, capable of selecting and transferring human poses that compose the animation of 2D characters. This process involved tracking artist poses from input video, clustering them, and using the clusters as a reference to create a character. The algorithm then automatically drives animation using pose data from a new video with different scenes. The authors demonstrate the effectiveness of their approach through qualitative feedback from artists, experiments, and comparisons with other animation techniques.

In addition to transferring them to an avatar, poses performed by an individual have various other applications, such as assisting in surveillance cameras for monitoring dangerous situations [5] or in detecting the risk of patient falls at the bedside [6].

Another example is proposed by [7], where a novel methodology based on computer vision and machine learning for remote tracking of patients' body joints during physiotherapeutic rehabilitation exercises was introduced. The system comprises two architectures, one capable of identifying the exercise being performed based on the patient's pose and another capable of measuring exercise correctness if performed incorrectly. Both modules of the architecture achieved over 90% accuracy in exercise recognition and validation.

Poses also have a significant impact on public speaking to large audiences in lectures or congresses. Understanding how world-renowned speakers behave in their presentations can benefit not only aspiring speakers but also professionals seeking to enhance their communication skills.

To aid in this process, the work by [8] proposed a visual analysis system that allows for the exploration and analysis of verbal and non-verbal presentation techniques in lectures, providing insights into their temporal distribution, co-occurrences, and contextualized exploration of individual videos. A case study involving experts in language education and university students provided anecdotal evidence of the approach's effectiveness and reported new findings on lecture presentation techniques. Quantitative feedback from a user study confirmed the utility of the visual system for multimodal analysis of video collections.

Thus, the present study aims to contribute to the field of motion analysis and posture patterns by offering new insights into how speakers and presenters behave in front of an audience during events and presentations. To achieve this, a dataset consisting of TED Talks-sponsored lectures was generated for the utilization of the pose identification algorithm, Openpifpaf. From these poses, the K-means clustering algorithm was applied to categorize them into common groups that we called keyposes. Finally, for the validation of the generated keyposes, an online questionnaire was implemented, with the participation of 35 respondents.

The main contribution of this paper are as follows:

- Creation of a dataset of poses during a lecture.
- Tool for the analysis of variability in body language during a speech or presentation.

- Proposal of a set of keyposes (pose alphabet) for representing body language in speech.
- Tool for transcribing, in a human-readable language, the sequence of movements performed during a speech.
- Tool to assist in detecting biases or redundancies in behavior during a speech.

The remainder of this paper is organized as follows: Section II outlines the methodology development process applied in this study. Section III presents the obtained results, and Section IV discusses the conclusions.

II. Materials and Methods

This section aims to present the proposed methodology and the dataset used in this work. Fig. 1 illustrates the main steps of the method proposed. In the first step, videos will be selected for pose recognition in the next step. Information from these poses will be extracted and used for clustering them using K-means. Finally, an evaluation of the generated groups is performed.

A. Dataset Generation

1) Video Selection: In this stage, a dataset comprised of lectures associated with the TED organization (Technology, Entertainment, and Design), covering a wide range of topics from science, technology, and business to art, culture, health, and education, was utilized. TED Talks are released under a Creative Commons BY-NC-ND license for unrestricted use. The selected videos were part of the playlist ”The most popular TED Talks of all time”[1], which features the top 25 most-viewed TED Talks of all time. From this playlist, a filtering process was carried out, resulting in 16 videos. The exclusion criteria involved removing lectures in which the speaker was using a handheld microphone, standing behind a podium, sitting, or using a flipchart.

2) Pose Recognition: This stage involves collecting frames from the dataset generated in the previous step and estimating the joints of individuals in each frame. For this purpose, the OpenPifPaf algorithm, proposed by [9], was employed. It is a real-time capable framework for detecting human and animal body joints. The program's output is a set of keypoints (x, y, c) containing pixel coordinates (x, y) and a confidence score c for each joint for each person in the respective video frame. From this set of keypoints, a data cleaning process was conducted to remove duplicated data or those with incomplete joints ($c \leq 0$).

Finally, two datasets were generated, one with the set of incomplete keypoints and another with the set of complete keypoints, both containing 34 attributes representing the x and y coordinates of each of the 17 keypoints returned by the algorithm. In this work, only the dataset with complete keypoints were used.

B. Proposed System

1) Feature Extraction: From the set of estimated points, a data transformation was performed to calculate the angles between the joints. Fig. 2 illustrates a scheme for angle calculation based on the coordinates obtained in the previous stage.

Given the vectors $\mathbf{u}$ and $\mathbf{v}$, obtained from points $P1$, $P2$, and $P3$, it is possible to calculate the dot product between them using Equation 1.

$$\mathbf{u} \cdot \mathbf{v} = \|\mathbf{u}\| \cdot \|\mathbf{v}\| \cdot \cos(\theta) \tag{1}$$

Rearranging Equation 1, the angle between the vectors can be found using Equation 2.

$$\theta = \arccos\left(\frac{\mathbf{u} \cdot \mathbf{v}}{\|\mathbf{u}\| \cdot \|\mathbf{v}\|}\right) \tag{2}$$

As a result, a new dataset containing six attributes was created, with each attribute representing an angle formed by certain joints among the 17 keypoints found in the original dataset.

2) Clustering: After the data transformation process, the clustering process begins. In this stage, the unsupervised learning algorithm K-means, developed by [10], will be used to group and identify different poses that an individual adopts during a lecture. However, since the value of k is not known, it is necessary to employ techniques to assist in choosing this value. In this project, the techniques used include the Silhouette Score [11], the Calinski-Harabasz Index [12], and the Elbow Method [13].

C. Cluster Evaluation

As a means of validating the generated pose groups from the previous stage, an online questionnaire was created using the Google Forms platform to collect people's opinions on the clustering performed by the unsupervised algorithm. This method can provide insight into how well the clusters represent similar poses in images based on each individual's interpretation and criteria.

The questionnaire consisted of multiple-choice questions in which a randomly selected image corresponding to a TED Talk was presented alongside two options of poses. It was the participant's task to choose which pose resembled the displayed image more. Fig. 3 illustrates an example of a question presented to the participants. Each question had a correct answer, and each cluster was compared to all the others.

III. Results and Discussion

A. Resulting dataset

The data cleaning process described in Section II-A2 resulted in a dataset with 26.469 pose with complete joints and another with 583.395 poses with incomplete joints. As mentioned earlier, only the dataset with complete poses were used in this work. Fig. 4 illustrates the pose generated from the average of all joints of the used dataset.

[1] https://www.ted.com/playlists/171/the_most_popular_ted_talks_of_all_time

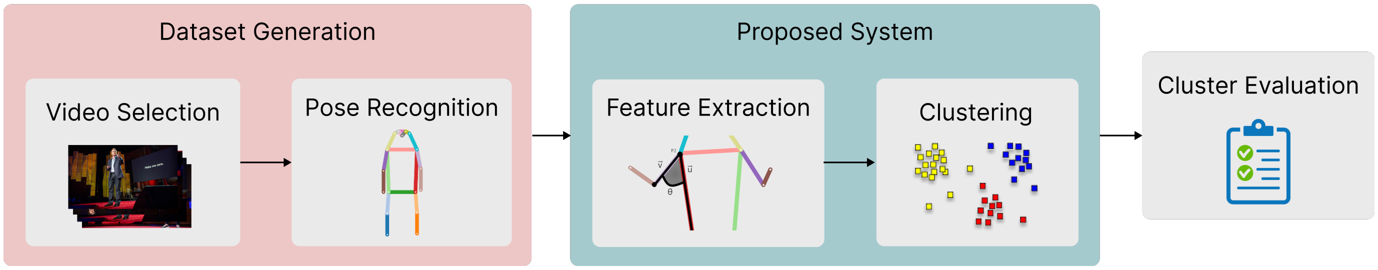


Fig. 1: Pipeline of the proposed methodology.

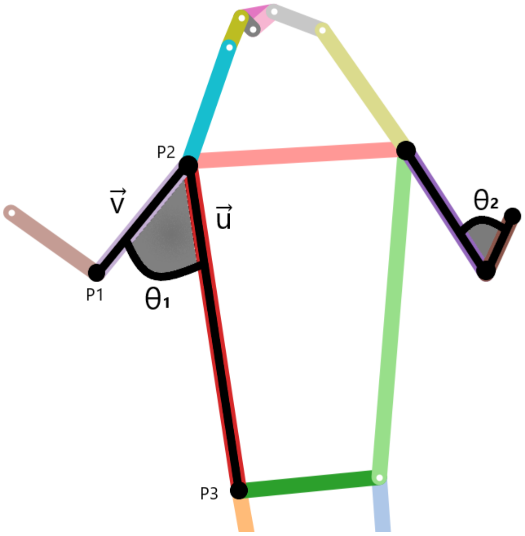


Fig. 2: Joint angle calculation.

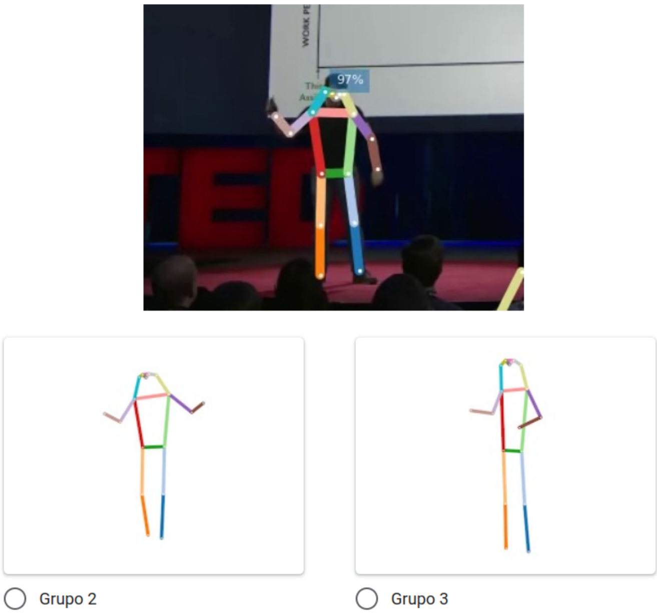


Fig. 3: Questionnaire for Cluster Validation

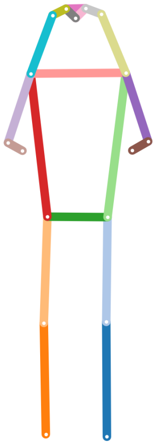


Fig. 4: Average of the poses in the dataset.

B. *Estimating optimal value of K*

A value of $k = 4$ was chosen for the number of clusters for the K-means algorithm. This choice was based on the elbow method, as depicted in Fig. 5, as well as the silhouette index and the Calinski-Harabasz coefficient, as shown in Fig. 6.

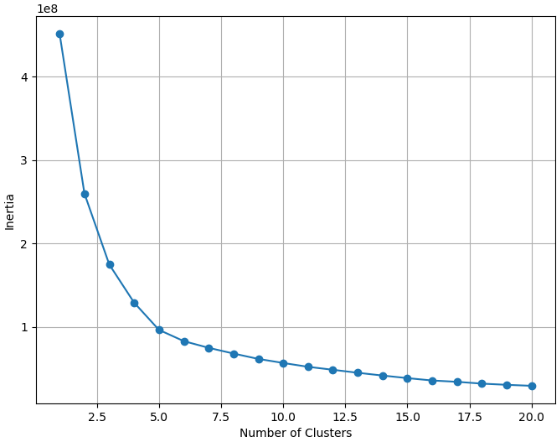


Fig. 5: Elbow method

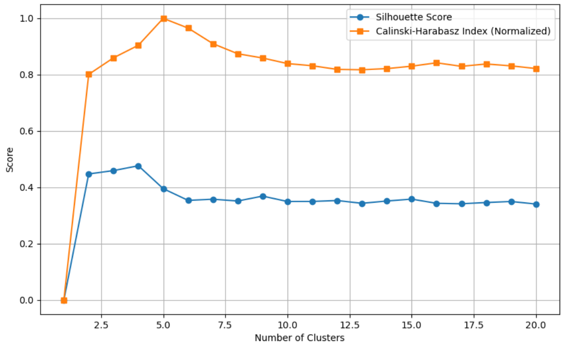
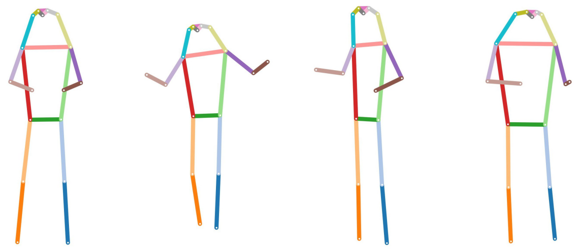


Fig. 6: Calinski-Harabasz Index and Silhouette Score

C. Centroid Estimation

Since the centroids are points calculated by the algorithm during the clustering process and do not exist in the original dataset, it is not possible to reconstruct a pose for visualization. This is because there was a transformation $\mathbb{R}^{34} \rightarrow \mathbb{R}^{6}$ of the samples, as described in Section II-B1. Therefore, the adopted visualization approach involved selecting the sample closest to the centroid of each cluster, resulting in the keyposes presented in Fig. 7.

There is a significant concentration of data in the first cluster, as shown in Fig. 8. This justifies the similarity between the mean pose of the entire dataset and that of the first cluster (Figs. 4 and 7a, respectively).



(a) Cluster 1 (b) Cluster 2 (c) Cluster 3 (d) Cluster 4

Fig. 7: Closest sample to the centroid of each cluster.

D. Questionnaire Evaluation

The questionnaire received responses from 35 participants, consisting of undergraduate and postgraduate students from the FEI University Center.

The correct classification of poses resulted in an accuracy of 85.71%, representing how well the four keyposes generated by K-means correctly generalize the poses commonly performed in lectures. However, participants encountered difficulty in classifying the third cluster in questions where it was compared to the first or the second clusters, as demonstrated in the confusion matrix presented in Fig. 9

On the other hand, the second cluster obtained the highest number of correctly classified responses (104), indicating that poses with more open arms, as depicted in Fig. 7b, are easier to distinguish.

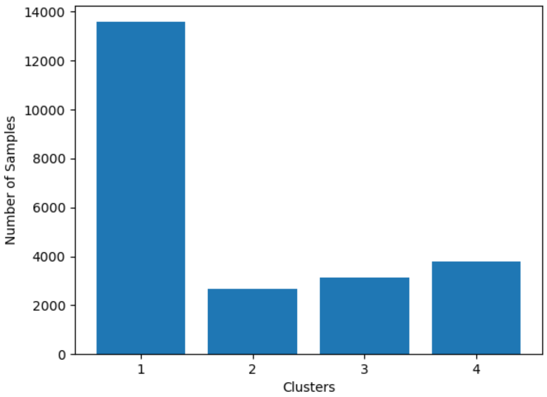


Fig. 8: Distribution of samples among clusters.

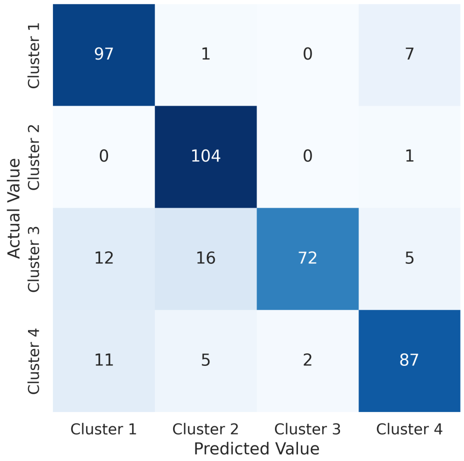


Fig. 9: Confusion Matrix.

In accordance with the confusion matrix, Fig. 10 displays the number of correct answers for each question and highlights an average accuracy of only 68.6% for correctly classified responses in Cluster 3. Question number eight, which had the lowest accuracy, is shown in Fig. 3.

IV. Conclusion

In this work, we proposed an analysis of posture patterns commonly adopted by speakers and presenters during lectures and conferences. Inertia, the Calinski-Harabasz coefficient, and silhouette index metrics indicated that an individual typically performs four to five different variations of poses during a presentation, with most of them involving arms close to the body. A questionnaire involving 34 individuals was conducted, asking 12 times if, given an input image, the keypose returned by the system better represents the image's pose compared to a randomly chosen key pose. The results indicate an 85.71% accuracy in choosing the pose returned by the system. As future work, techniques involving data imputation may be

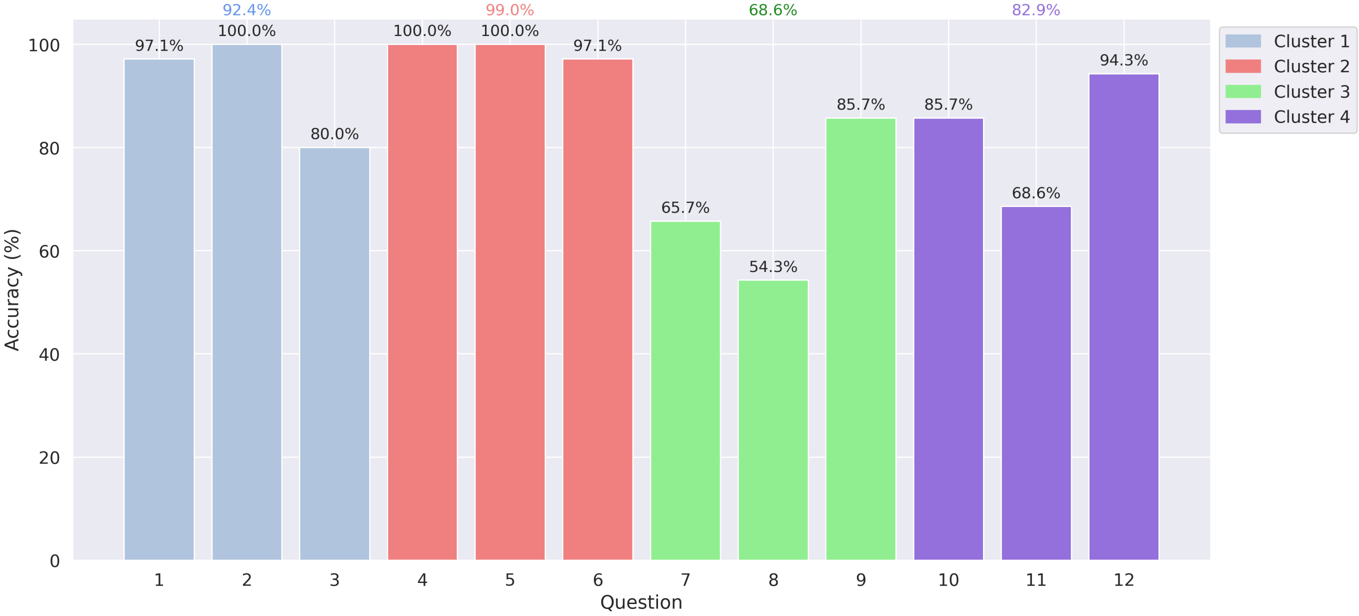


Fig. 10: Percentage of correct answers for each question in the questionnaire. Each color represents the correct cluster corresponding to each question.

relevant to increase the sample size in the dataset, given that many lectures have multiple cameras, and few of them feature full-body shots of the speaker.

Acknowledgment

The authors would like to thank the Centro Universitário FEI, CAPES, FAPESP, and CNPq for the support they have given to this work.

References

[1] Paradisi, P., Raglianti, M., and Sebastiani, L. (2021). Online Communication and Body Language. Frontiers in Behavioral Neuroscience, 15.

[2] Wang, I., and Ruiz, J. (2021). Examining the Use of Nonverbal Communication in Virtual Agents. International Journal of Human–Computer Interaction, 37, 1648 - 1673.

[3] Jeong, D.C., Xu, J.J., and Miller, L.C. (2020). Inverse Kinematics and Temporal Convolutional Networks for Sequential Pose Analysis in VR. 2020 IEEE International Conference on Artificial Intelligence and Virtual Reality (AIVR), 274-281.

[4] Willett, N.S., Shin, H.V., Jin, Z., Li, W., and Finkelstein, A. (2020). Pose2Pose: pose selection and transfer for 2D character animation. Proceedings of the 25th International Conference on Intelligent User Interfaces.

[5] Miki, D., Abe, S., Chen, S., and Demachi, K. (2020). Robust human pose estimation from distorted wide-angle images through iterative search of transformation parameters. Signal, Image and Video Processing, 14, 693-700.

[6] Qiu, R., Teng, W., Wei, Z., Zhang, C., Zhong, Y., and Zhai, J. (2022). Fall Detection Algorithm Based on Lightweight Openpose Model with Attention Mechanism. Academic Journal of Science and Technology.

[7] Francisco, J.A. and Rodrigues, P.S. (2022). Computer Vision Based on a Modular Neural Network for Automatic Assessment of Physical Therapy Rehabilitation Activities. IEEE Transactions on Neural Systems and Rehabilitation Engineering, 31, 2174-2183.

[8] Wu, A. and Qu, H. (2020). Multimodal Analysis of Video Collections: Visual Exploration of Presentation Techniques in TED Talks. IEEE Transactions on Visualization and Computer Graphics, 26, 2429-2442.

[9] Kreiss, S., Bertoni, L., and Alahi, A. (2021). OpenPifPaf: Composite Fields for Semantic Keypoint Detection and Spatio-Temporal Association. IEEE Transactions on Intelligent Transportation Systems, 23, 13498-13511.

[10] MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations.

[11] Rousseeuw, P.J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics, 20, 53-65.

[12] Caliński, T., and Harabasz, J. (1974). A dendrite method for cluster analysis. Communications in Statistics-theory and Methods, 3, 1-27.

[13] Thorndike, R.L. (1953). Who belongs in the family? Psychometrika, 18, 267-276.